Datasheets for Datasets

1 Links

We attach our web page for the NeurIPS 2021 Datasets and Benchmarks Track reviewers: **Dataset:** `https://bit.ly/38prlVO`

Our dataset with documentation and associated code will be maintained under the following license: CC-BY-4.0 (`https://creativecommons.org/licenses/by/4.0/`).

2 Author Statement

Our dataset can be commonly copied and redistributed in any medium or format. Our dataset can be accessed using the following license: CC-BY-4.0 (`https://creativecommons.org/licenses/by/4.0/`).

3 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Deepfakes have become a challenging well-known technical, social, and ethical issue now. Therefore, many ethical, security, and privacy concerns arise due to the ease of generating deepfakes. And, there is a strong need for developing good deepfake detection methods. However, a good deepfake dataset is required. Although there exisit deepfake datasets, there exist no multimodal (video+audio) deepfake dataset. In this work, we aim to fill the gap by providing multimodal (video+audio) deepfake dataset.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

Every author in the paper has contributed.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was supported by the Basic Science Research Program through National Research Foundation of Korea (NRF) grant funded by Korea government MSIT (No. 2020R1C1C1006004). Also, this research was partly supported by IITP grant funded by the Korea government MSIT (No. 2021-0-00017, Original Technology Development of Artificial Intelligence Industry).

4 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? How many instances are there in total (of each type, if appropriate)?

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

Our FakeAVCeleb consists of 20,000+ deepfake videos of 490 celebrities with different ethnic backgrounds belonging to diverse age groups. The real videos that are used to create deepfake dataset are taken from VoxCeleb2 dataset, which consists of real YouTube videos of 6,112 celebrities. . Each ethnic group have a separate directory, and videos are sorted with respect to gender. There are four categories of datasets, for real and three fake, each contains a separate directory.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? We include all possible instances or combinations to the best of our knowledge with respect to Audio-Video pairs (($\mathcal{A}_{\mathcal{R}}\mathcal{V}_{\mathcal{R}}$),($\mathcal{A}_{\mathcal{R}}\mathcal{V}_{\mathcal{F}}$),($\mathcal{A}_{\mathcal{F}}\mathcal{V}_{\mathcal{R}}$),($\mathcal{A}_{\mathcal{F}}\mathcal{V}_{\mathcal{F}}$)).

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The dataset consists of processed videos and audios.

Is there a label or target associated with each instance?

Yes. The directory structure of our dataset corresponds to the labels.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No, all of the annotation information has been provided.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Yes. The real videos and generated fake videos from are sorted and named explicitly to indicate its type. (e.g., REAL-A/MEN/id00220/iPfJcAJQ84Y/00027.mp4, FAKE-FACESWAP-C/MEN/id00220/00002-id00575-iPfJcAJQ84Y-faceswap-32/00027-fake.mp4)

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes, we provide *JSON* files representing each data split for reproducibility of the detection performance.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. We tried our best to filter out the bad samples and mistakes with several researchers independently cross-checking samples.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources,

Mostly, FakeAVCeleb is self-contained, however, since our real videos are from VoxCeleb2 and contains only 490 videos, researcher can refer to the VoxCeleb2 dataset if more real videos are required.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor patient confidentiality, data that includes the content of individuals' non-public communications)?

No, the dataset is generated real Youtube videos of celebrities. And it does not contain data that might be considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No, not at all.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. FakeAVCeleb related to people.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

Yes, the dataset is sorted and each file is named with respect to people with the specific ethnicity and gender.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?

Since the videos are of celebrities, it is possible to identify the people. However, there are already openly available dataset information from VoxCeleb2 and YouTube.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?

No. The dataset contains YouTube videos of celebrities with different ethic backgrounds. The intention to include videos of people with diverse ethnic background is to remove racial bias issues in dataset.

5 Collection

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

To create deepfake videos, we selected open sourced real YouTube videos from VoxCeleb2 dataset, which consists of real YouTube videos of 6,112 celebrities.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Three different researchers independently cross-checked the samples and we also used Face++ API to match the best real and fake pairs.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

No, not applicable.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

We used the dataset from VoxCeleb2, and three researchers worked on choosing the best samples according to different genders, races, etc, and independently verified the selections one another.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The dataset was created from Jan. to June 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No ethical review process was needed, because all datasets are already available on the Internet.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.

Yes. It is related to people.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

We obtained from VoxCeleb2, online data source.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

NA.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

NA.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

NA.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

NA.

6 Preprocessing, cleaning and labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Preprocessing was performed separately for videos and audios. Since we collected videos from VoxCeleb2 dataset, these videos are already face-centered and cropped. We extract respective frames from each video and store them separately, and then extract audios from the videos and store them in *.wav* format with a sampling rate of 16 kHz. Before inputting audio directly to the model for training, we first compute Mel-Frequency Cepstral Coefficients (MFCC) features by applying a $25ms$ Hann window [**?**] with $10ms$ window shifts, followed by a fast Fourier transform (FFT) with 512 points. As a result, we obtain a 2D array of 80 MFCC features $(D = 80)$ per audio frame and store the resulting MFCC features as a three channel image, which is then passed to the model as an input to extract speech features so that it learns the difference between real and fake human speeches.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw images are available.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Yes, we use Python packages to process our dataset.

7 Use Cases

Has the dataset been used for any tasks already? If so, please provide a description.

No.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. What (other) tasks could the dataset be used for?

We provide the information in the following URL: `http://doi.org/10.23056/FAKEAVCELEB_DASHLAB`

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

It is explained in our DOI site mentioned above.

Are there tasks for which the dataset should not be used? If so, please provide a description.

No, not applicable.

8 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Use of the dataset is free to all researchers after signing a data use agreement.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

We will provide the dataset through Google drive after receiving the Data Use Agreement (DUA) by Google form. Also, we have DOI: `http://doi.org/10.23056/FAKEAVCELEB_DASHLAB`

When will the dataset be distributed?

Our dataset is available: `http://doi.org/10.23056/FAKEAVCELEB_DASHLAB`.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Yes. it will be distributed under the following license: Attribution 4.0 International (CC BY 4.0) **CC BY.** This license allows requesters to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

NA.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

NA.

9 Maintenance

Who is supporting/hosting/maintaining the dataset? How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The authors of the paper are maintaining the dataset, particularly by Hasam Khalid: hasam.khalids@g.skku.edu.

Is there an erratum? If so, please provide a link or other access point.

No, not applicable.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, if there are issues with our dataset, we will immediately reflect them on our dataset site. Also, we keep and provide the updated version with the older version kept around for consistency.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

There are no applicable limits on the retention of the dataset.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes, each older version of the dataset will continue to be hosted on FakeAVCeleb dataset site: `http://doi.org/10.23056/FAKEAVCELEB_DASHLAB`. We add and manage all released versions of our dataset for easy access.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Others can contact us about incorporating fixes and extensions. We will be happy that our work can be extended further.